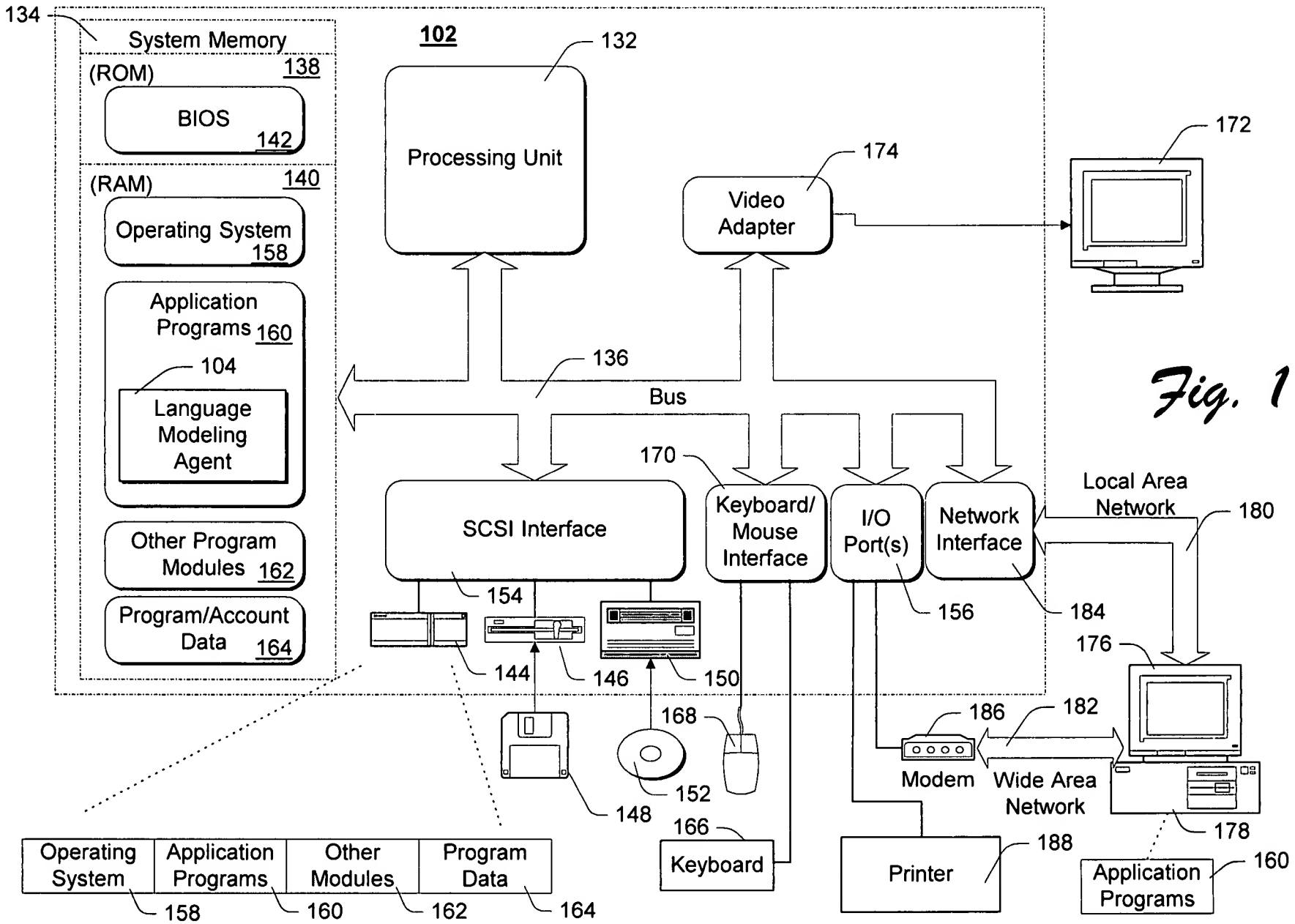


Fig. 1

MS1-441US

090590 09/709960

Fig. 2

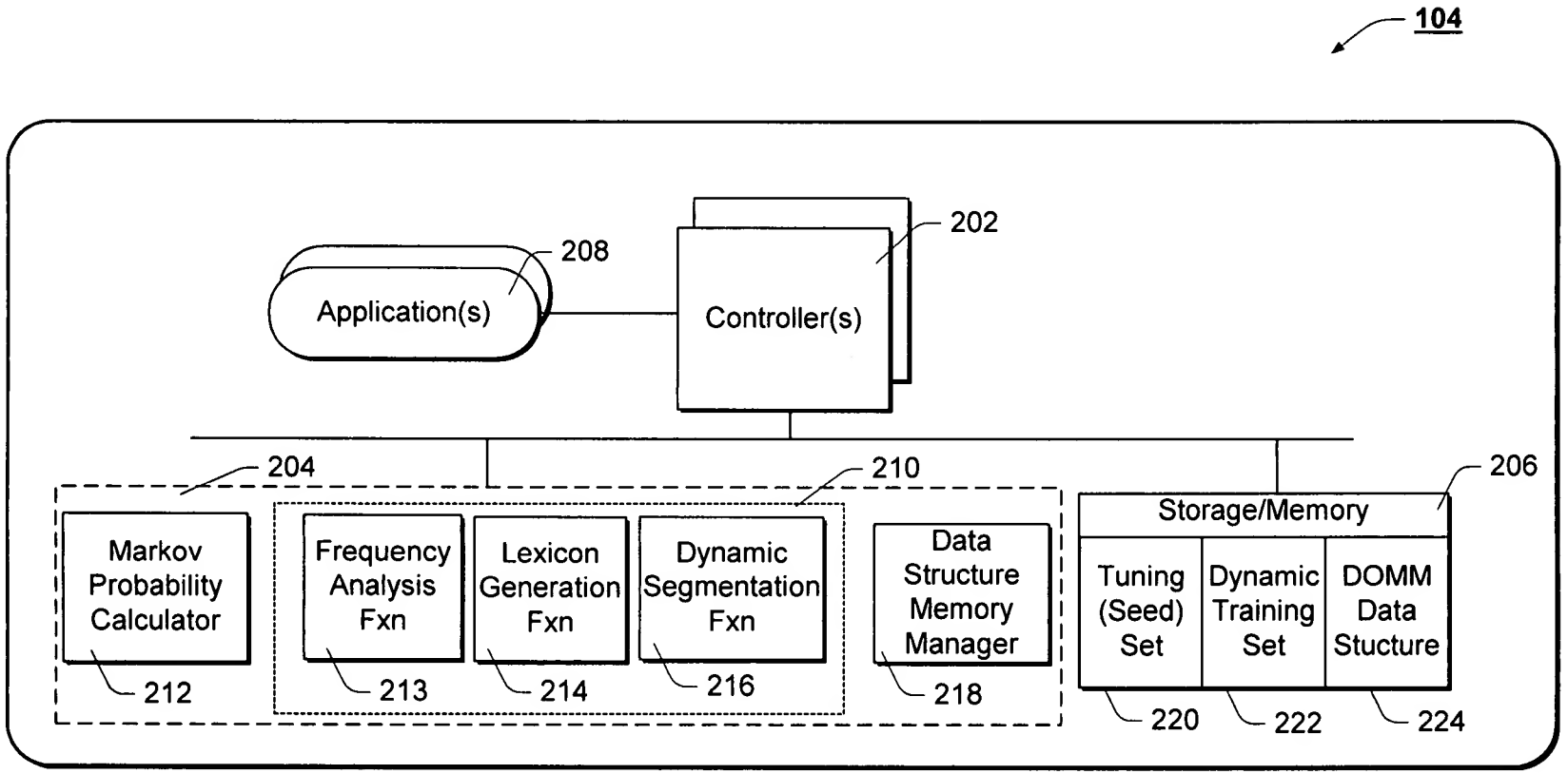


Fig. 3

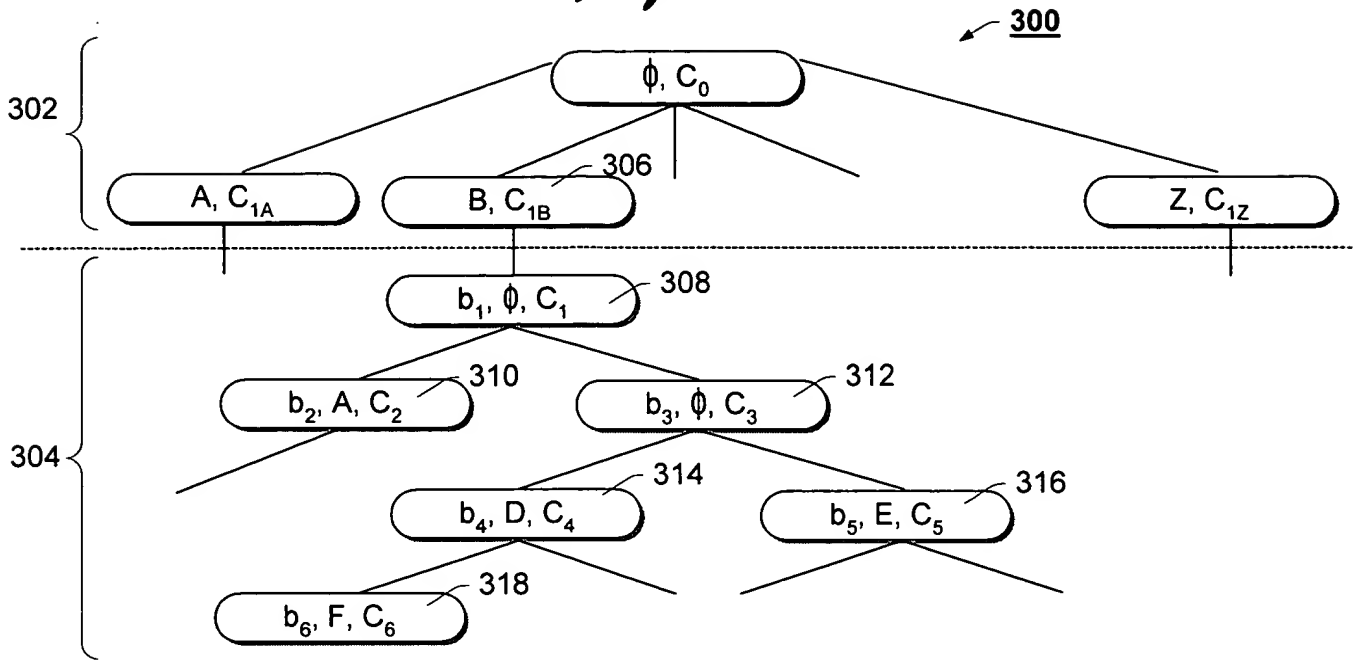

300
302
304
ϕ, C0
A, C1A
B, C1B
306
Z, C1Z
b1, ϕ, C1
308
b2, A, C2
310
b3, ϕ, C3
312
b4, D, C4
314
b5, E, C5
316
b6, F, C6
318

Fig. 4

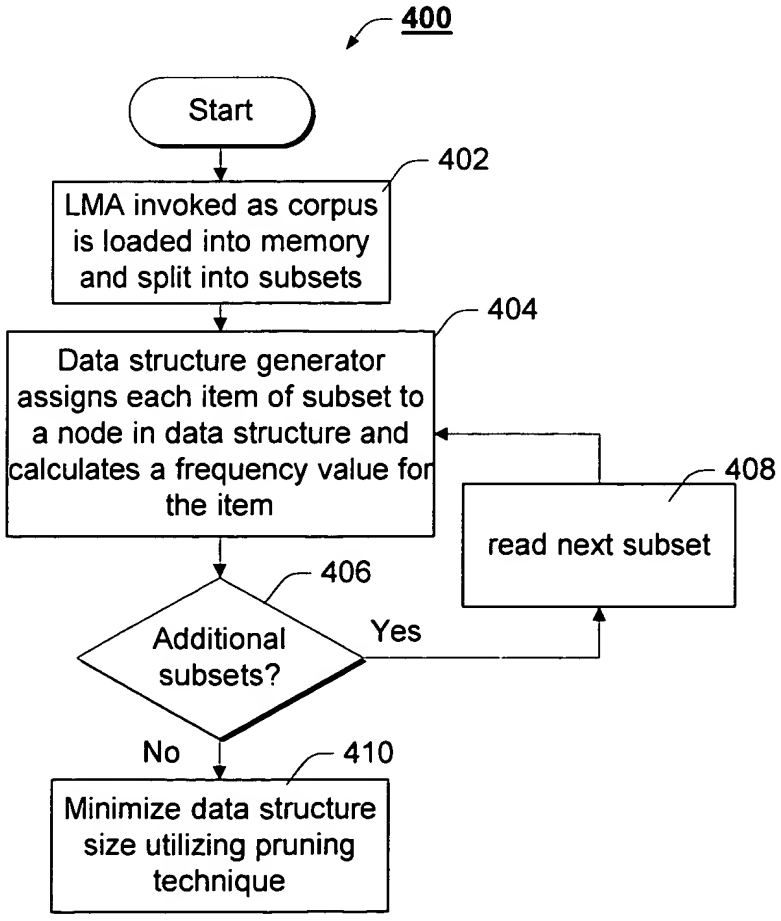

400
Start
LMA invoked as corpus is loaded into memory and split into subsets
402
Data structure generator assigns each item of subset to a node in data structure and calculates a frequency value for the item
404
read next subset
408
Additional subsets?
406
Yes
No
Minimize data structure size utilizing pruning technique
410

Fig. 5

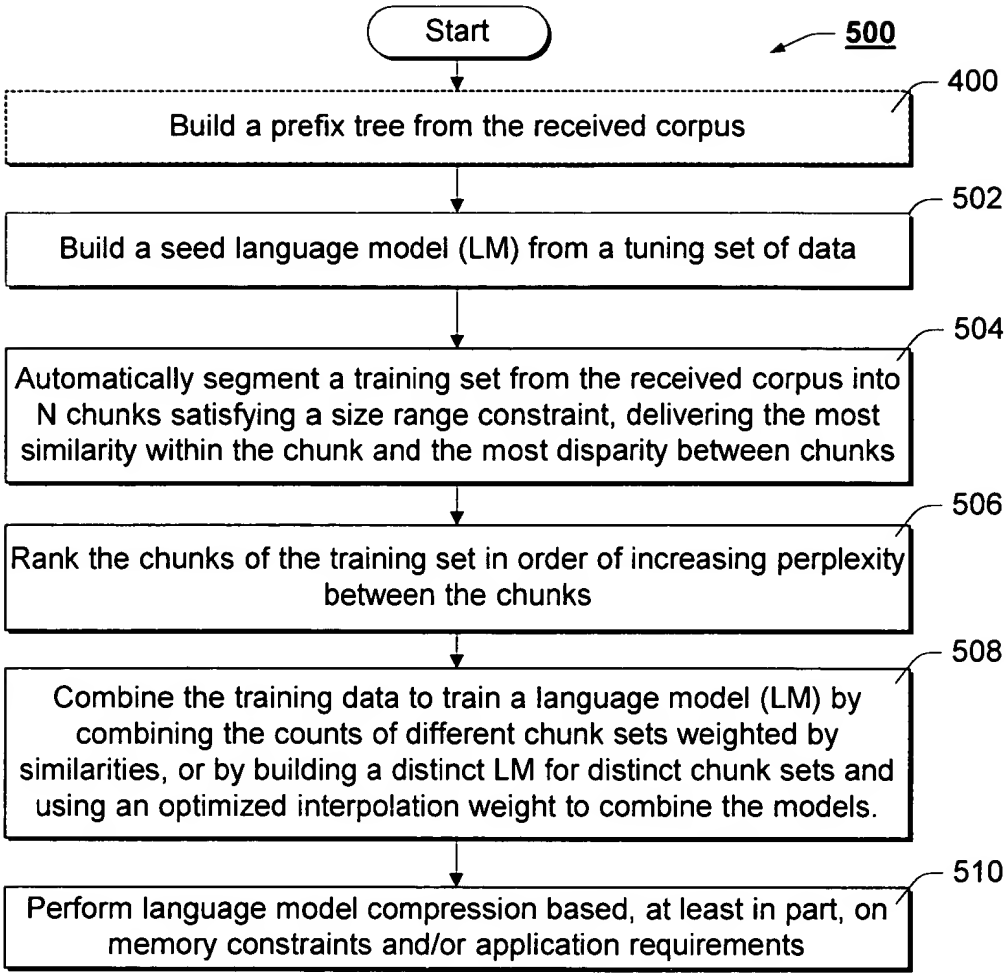

Start
500
400
Build a prefix tree from the received corpus
502
Build a seed language model (LM) from a tuning set of data
504
Automatically segment a training set from the received corpus into N chunks satisfying a size range constraint, delivering the most similarity within the chunk and the most disparity between chunks
506
Rank the chunks of the training set in order of increasing perplexity between the chunks
508
Combine the training data to train a language model (LM) by combining the counts of different chunk sets weighted by similarities, or by building a distinct LM for distinct chunk sets and using an optimized interpolation weight to combine the models.
510
Perform language model compression based, at least in part, on memory constraints and/or application requirements

Fig. 9

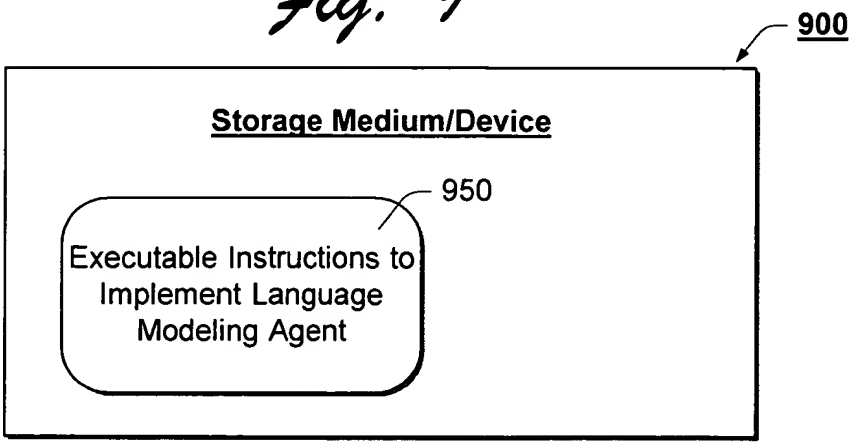
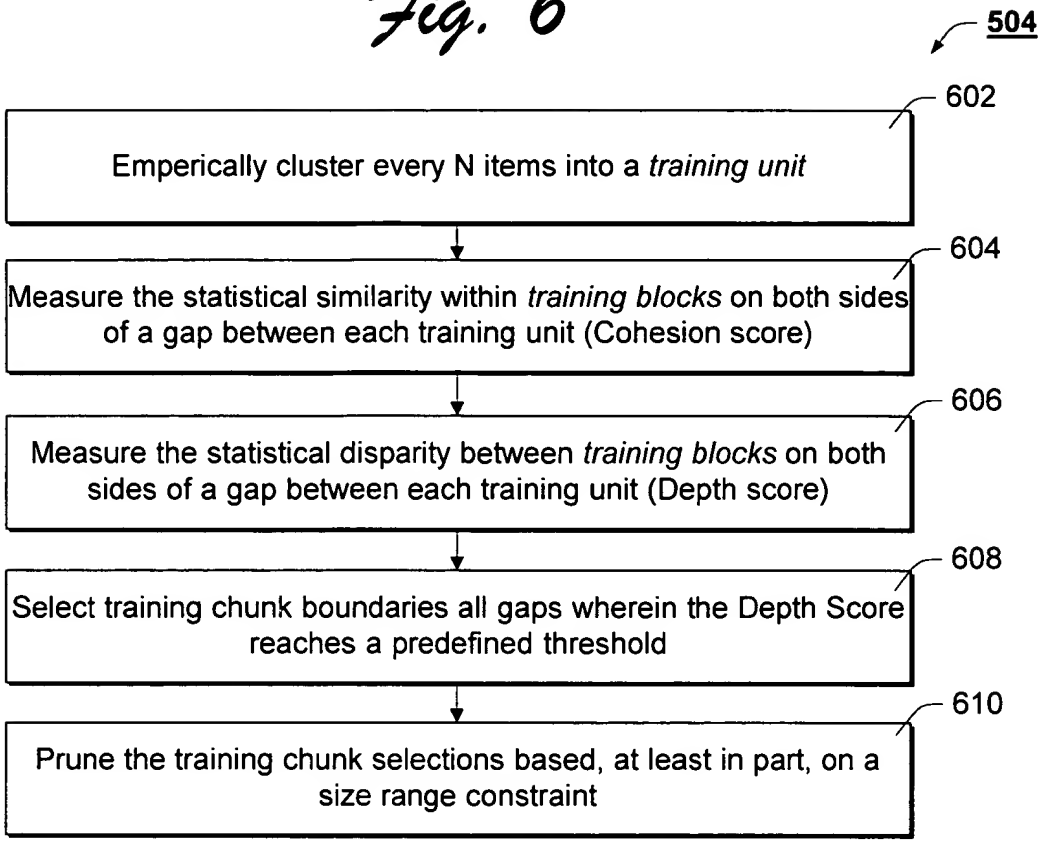

900
Storage Medium/Device
950
Executable Instructions to Implement Language Modeling Agent

Fig. 6



5605760.doc

Fig. 7

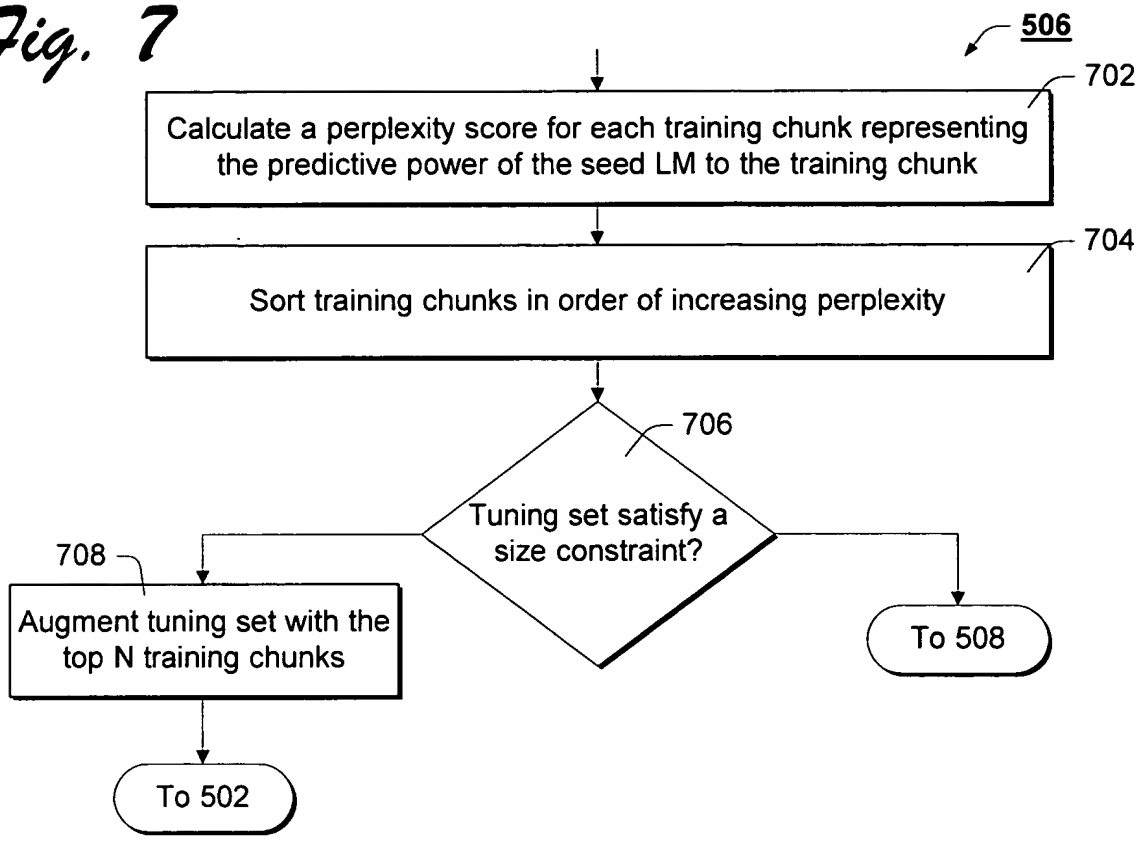
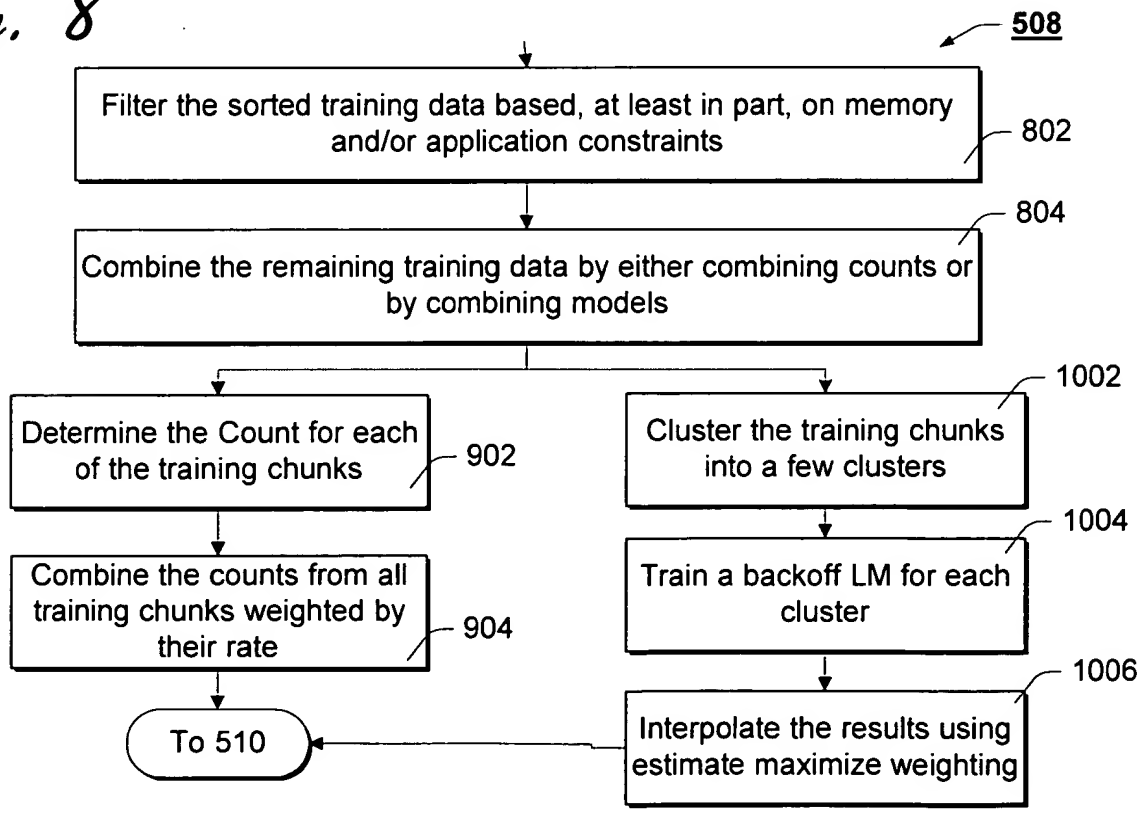

506
Calculate a perplexity score for each training chunk representing the predictive power of the seed LM to the training chunk
702
Sort training chunks in order of increasing perplexity
704
Tuning set satisfy a size constraint?
706
Augment tuning set with the top N training chunks
708
To 502
To 508

Fig. 8


508
Filter the sorted training data based, at least in part, on memory and/or application constraints
802
Combine the remaining training data by either combining counts or by combining models
804
Determine the Count for each of the training chunks
902
Combine the counts from all training chunks weighted by their rate
904
To 510
Cluster the training chunks into a few clusters
1002
Train a backoff LM for each cluster
1004
Interpolate the results using estimate maximize weighting
1006